Обучение машинного перевода без параллельных текстов*

Артеменков А. А. Гончаров М. Ю. Ярошенко А. М. Иванов А. В. Мазуров М. М. Борисова А. В. Скиднов Е. А. Строганов А. А. Рябов Ф. А.
Московский физико-технический институт

В данной работе исследуется задача машинного перевода между двумя языками. Для решения часто используются параллельные предложения, то есть совпадающие по смыслу фразы на двух языках. В работе рассматривается альтернативная модель, не требующая большого количества параллельных предложений. Она использует нейронную сеть типа Seq2Seq, имеющую скрытое пространство. Для проверки качества модели проводится вычислительный эксперимент по переводу предложений между близкими языками, такими как русский и украинский.

Ключевые слова: *нейронные сети, машинный перевод, автокодировщики.*

Введение

В зависимости от специфики пары языков выделяют несколько подходов к машинному переводу. При наличии достаточного числа параллельных предложений(порядка миллиона) использование глубоких нейронных сетей привело к получению хороших результатов [1],[2].

Но для многих пар языков нет достаточной базы примеров. Одним из подходов на основе параллельных предложений является пополнение обучающей выборки переводами с предыдущих итераций работы нейронной сети [3].

Ниже представлено решение задачи машинного перевода при отсутствии достаточного количества параллельных предложений [4], [5], [6]. В модели используются 2 типа автокодировщиков: рекуррентные нейронные сети [7], [8], которые реализуют перевод слов в скрытое пространство, и сеть-дискриминатор, определяющая по векторному представлению язык исходного предложения. Сети-энкодеры оптимизируются так, чтобы представление одного и того же предложения на разных языках совпадало в скрытом пространстве, то есть, чтобы дискриминатору было сложнее определить язык, к которому относится вектор. Обучение состоит из двух фаз. На первой оптимизируется работа дискриминатора: предложение кодируется с добавлением шума [9] и подаётся на вход и происходит перераспределение параметров. На второй стадии происходит перераспределение парамтеров уже у сетей-кодировщиков. После проведение этих шагов вычисляется значение функции потерь.

Такой подход был использован в [10] для пары языков французский-английский. В данной работе будет проведен схожий эксперимент для перевода с русского на украинский. Качество переводчика в работе оценивается с помощью метрики BLEU [11].

Постановка задачи

Даны обучающая выборка $\mathfrak{D}^{\text{src}} = \{\mathbf{s}_1^{\text{src}}, \ldots, \mathbf{s}_{m_{\text{src}}}^{\text{src}}\}$, $\mathfrak{D}^{\text{tgt}} = \{\mathbf{s}_1^{\text{tgt}}, \ldots, \mathbf{s}_{m_{\text{tgt}}}^{\text{tgt}}\}$, которая состоит из двух корпусов произвольных предложений для каждого языка; и валидационная

Научный руководитель: Стрижов В. В. Авторы: А.В. Грабовой, О.Ю. Бахтеев, В.В. Стрижов, Eric Gaussier, координатор Малиновский Г. С. Консультант: Бахтеев О. Ю.

выборка $\mathfrak{D}^{valid} = \{(\tilde{\mathbf{s}}_1^{\text{src}}, \tilde{\mathbf{s}}_1^{\text{tgt}}), \ldots, (\tilde{\mathbf{s}}_{m_{\text{valid}}}^{\text{src}}, \tilde{\mathbf{s}}_{m_{\text{valid}}}^{\text{tgt}})\}$, которая состоит из пар параллельных предложений.

Задача состоит в том, чтобы, используя обучающую выборку, построить модель M перевода предложений с языка src на язык tgt и выбрать ее параметры таким образом, чтобы максимизировать среднее значение метрики BLEU на валидационной выборке

$$\frac{1}{m_{\text{valid}}} \sum_{i=1}^{m_{\text{valid}}} \text{BLEU}(M(\tilde{\mathbf{s}}_i^{\text{src}}), \tilde{\mathbf{s}}_i^{\text{tgt}}) \to \max_M$$

Гипотеза

Пусть $\mathcal{S}^l$ – множество всех предложений на языке $l \in \{\text{src}, \text{tgt}\}$. Выдвигается гипотеза о том, что существует единое латентное пространство $\mathcal{L}$ и отображения $\boldsymbol{f}^l : \mathcal{S}^l \to \mathcal{L}$ и $\boldsymbol{g}^l : \mathcal{L} \to \mathcal{S}^l$ такие, что

— для $\forall l_1, l_2 \in \{\text{src}, \text{tgt}\}$ и $\forall s^{l_1} \in \mathcal{S}^{l_1}$ $\boldsymbol{g}^{l_2}(\boldsymbol{f}^{l_1}(s^{l_1}))$ совпадает с s^{l_1}, если l_1 совпадает с l_2, и является корректным переводом s^{l_1}, если l_1 и l_2 различаются
— распределения образов $\boldsymbol{f}^{\text{src}}(\mathcal{S}^{\text{src}})$ и $\boldsymbol{f}^{\text{tgt}}(\mathcal{S}^{\text{tgt}})$ совпадают

Описание метода

Предлагаемый метод заключается в том, чтобы для каждого $l \in \{\text{src}, \text{tgt}\}$ моделировать отображения $\boldsymbol{f}^l$ и $\boldsymbol{g}^l$ кодировщиком $\mathbf{f}^l$ и декодировщиком $\mathbf{g}^l$ соответственно. Таким образом, моделью перевода является композиция $\mathbf{g}^{\text{tgt}} \circ \mathbf{f}^{\text{src}}$.

Оптимизация проводится следующим образом. Функция ошибки содержит три слагаемых, которые соответствуют сделанным предположениям об отображениях $\boldsymbol{f}^l$ и $\boldsymbol{g}^l$.

Ошибка восстановления Для каждого $l \in \{\text{src}, \text{tgt}\}$ рассматривается входное предложение $\mathbf{s}^l$ на языке l. Ошибки считаются между $\mathbf{s}^l$ и его образом при отображении $\mathbf{g}^l \circ \mathbf{f}^l$.

Ошибка перевода Рассматривается предложение $\mathbf{s}^{\text{src}}$ без ограничения общности на языке src. В качестве входного предложения используется $\mathbf{s}^{\text{tgt}}$ – перевод $\mathbf{s}^{\text{src}}$, полученный с помощью некоторой слабой модели перевода M^0. Ошибка перевода с языка tgt на язык src считается между $\mathbf{s}^{\text{src}}$ и образом $\mathbf{s}^{\text{tgt}}$ при отображении $\mathbf{g}^{\text{src}} \circ \mathbf{f}^{\text{tgt}}$. Аналогичным образом считается ошибка перевода с языка src на язык tgt. В качестве M^0 используется пословный переводчик на основе предобученных векторных представлений слов $\mathcal{E}^{\text{src}} = \{\mathbf{x}_i\}_{i=1}^{n_{\text{src}}}$ и $\mathcal{E}^{tgt} = \{\mathbf{y}_i\}_{i=1}^{n_{\text{tgt}}}$. Переводом слова $\mathbf{x} \in \mathcal{E}^{\text{src}}$ является $\mathbf{y} = \arg\min_{\mathbf{y} \in \mathcal{E}^{\text{tgt}}} \rho(\mathbf{x}, \mathbf{y})$, где ρ – косинусное расстояние (перевод слов языка tgt осуществляется аналогично).

Штраф за различие распределений Вводится дискриминатор $\mathbf{d}$, который решает задачу классификации векторов $\mathbf{h}^l$ латентного пространства $\mathcal{L}$ на классы $\{0, 1\}$: $\mathbf{h}^l \in 0 \Leftrightarrow l = \text{src}$. Векторы $\mathbf{h}^l$ получаются кодировщиком $\mathbf{f}^l$, в качестве промежуточного результата при отображениях входных предложений. В функцию ошибки добавляется штраф за то, что дискриминатор точно определяет язык входного предложения. Таким образом параметры модели оптимизируются таким образом, чтобы усложнить задачу дискриминатору. В свою очередь параметры дискриминатора оптимизируются параллельно параметрам модели. Соревновательный процесс оптимизации мотивирован желанием добиться сходства распределений латентных векторов для разных языков.

Детали метода

Введем словарь V^{src}, содержащий проиндексированные слова, встречающиеся в предложениях из $\mathfrak{D}^{\text{src}}$ и переводах предложений из $\mathfrak{D}^{\text{tgt}}$, получаемых с помощью M^0. Аналогично введем V^{tgt}.

Предложения на разных языках, использующиеся при оптимизации кодируются с помощью соответствующих словарей. Входные предложения зашумляются преобразованием σ: сначала с некоторой вероятностью q из них удаляется каждое слово, а затем производится случайная перестановка оставшихся слов с условием, что слово не может оказаться дальше чем на k позиций от своей начальной позиции (q, k – гиперпараметры).

Кодировщик $\mathbf{f}^l$ включает в себя слой векторного представления слов $\mathbf{e}^l$, параметры которого инициализируются векторами $\mathcal{E}^l$, и реккурентную нейронную сеть $\mathbf{r}^{\text{enc}}$. Декодировщик $\mathbf{g}^l$ включает в себя $\mathbf{e}^l$, $\mathbf{r}^{\text{dec}}$ и классификатор $\mathbf{c}^l$, который решает задачу многоклассовой классификации выходов $\mathbf{r}^{\text{dec}}$ на классы, соотвествующие словам в словаре V^l (отображает выходы $\mathbf{r}^{\text{dec}}$ в векторы вероятностей размерности $|V^l|$). Параметры $\mathbf{r}^{\text{enc}}$ и $\mathbf{r}^{\text{dec}}$ общие для $\mathbf{f}^{\text{src}}$, $\mathbf{f}^{\text{tgt}}$ и $\mathbf{g}^{\text{src}}$, $\mathbf{g}^{\text{tgt}}$ соответственно.

В качестве меры ошибок классификации используется кросс-энтропия CE. Итоговый вид функции ошибки

$$L_{\text{tran}} = w_1 \cdot \sum_{l \in \{src,tgt\}} CE(\mathbf{g}^l(\mathbf{f}^l(\sigma(\mathbf{s}^l))), \mathbf{s}^l) +$$

$$+ w_2 \cdot \sum_{l_1 \neq l_2 \in \{src,tgt\}} CE(\mathbf{g}^{l_2}(\mathbf{f}^{l_1}(\sigma(\mathbf{s}^{l_1}))), \mathbf{s}^{l2}) + w_3 \cdot \sum_{l \in \{src,tgt\}} CE(\mathbf{d}(\mathbf{f}^l(\sigma(\mathbf{s}^l))), \mathbb{I}[l = \text{src}])$$

$$L_{\text{disc}} = \sum_{l \in \{src,tgt\}} CE(\mathbf{d}(\mathbf{f}^l(\sigma(\mathbf{s}^l))), \mathbb{I}[l = \text{tgt}]).$$

Результаты экспериментов

В качестве параллельного корпуса была взята выборка на основе OpenSubtitles-2018, в качестве непараллельной выборки - мультиязычные статьи из википедии. Работа была проделана на англо-французской выборке, так как качество найденных русско-украинских выборок не является достаточным как для построения хорошей модели, так и для валидации: в предложениях присутствует слишком много слов не из словаря (числа и формы слов), а сами переводы не всегда точны. Ниже представлены графики зависимости BLEU от номера итерации при различных размерах обучающей выборки и различной размерности скрытого пространства $\mathcal{L}$. Можно сделать вывод, что из-за увеличения числа предложений в обучающей выборке BLEU растёт более медленно, но при этом мы теряем некоторую часть информации о распределении предложений, и ограничиваем максимально достижимое качество. Кроме того, при увеличении размера скрытого пространства модель оптимизируется быстрее, однако при этом повышается риск переобучения.

Вывод

В данной работе была исследована задача машинного перевода между двумя языками. Предложен подход, основанный на моделях автокодировщиков и не требующий наличия большого корпуса параллельных предложений. Возможна дальнейшая работа по модификации алгоритма таким образом, чтобы это не так сильно влияло на качество перевода [12], [13]. В данных работах предлагается извлечь дополнительную информацию из выборки для одного языка, что позволит сохранить независимость метода от наличия параллельных предложений.

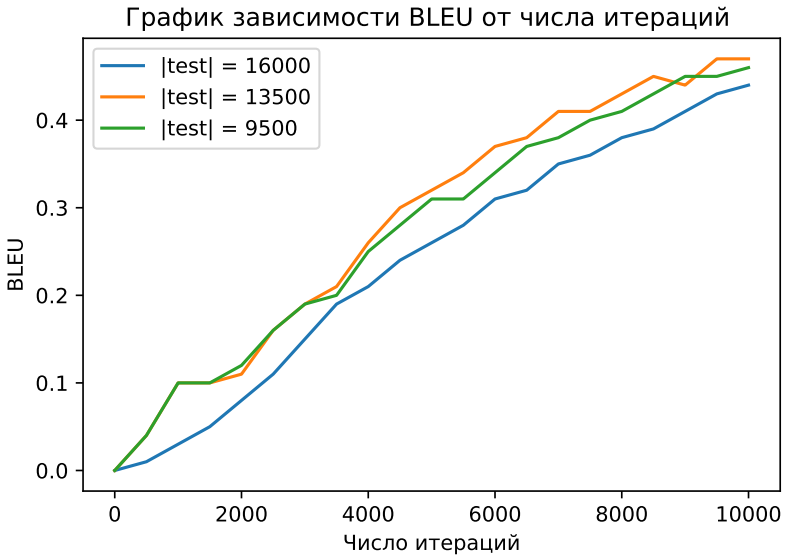

Рис. 1. Зависимость BLEU от размера обучающей выборки.

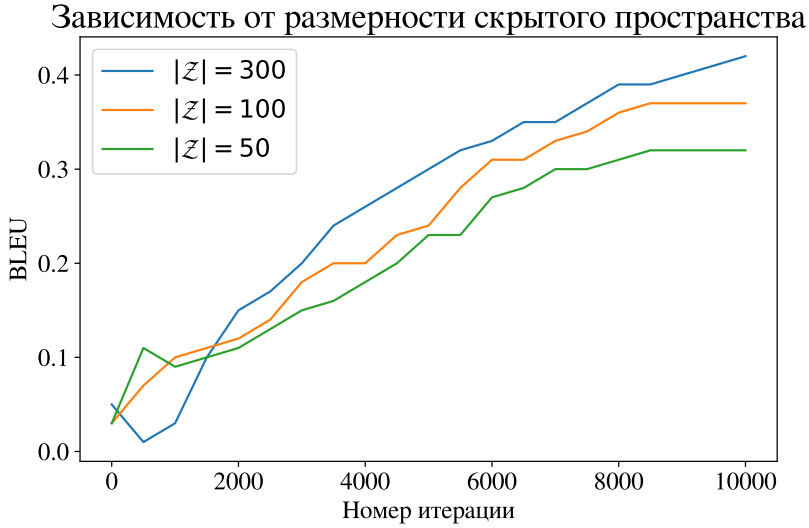

Рис. 2. Зависимость BLEU от размерности скрытого пространства.

Литература

[1] Bilingual word embeddings for phrase-based machine translation / Will Y Zou, Richard Socher, Daniel Cer [и др.] // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. С. 1393–1398.

[2] On the properties of neural machine translation: Encoder-decoder approaches / Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau [и др.] // arXiv preprint arXiv:1409.1259. 2014.

[3] Bertoldi Nicola, Federico Marcello. Domain adaptation for statistical machine translation with monolingual resources // Proceedings of the fourth workshop on statistical machine translation / Association for Computational Linguistics. 2009. С. 182–189.

[4] Google's neural machine translation system: Bridging the gap between human and machine translation / Yonghui Wu, Mike Schuster, Zhifeng Chen [и др.] // arXiv preprint arXiv:1609.08144. 2016.

[5] Sutskever Ilya, Vinyals Oriol, Le Quoc V. Sequence to sequence learning with neural networks // Advances in neural information processing systems. 2014. С. 3104–3112.

[6] Bahdanau Dzmitry, Cho Kyunghyun, Bengio Yoshua. Neural machine translation by jointly learning to align and translate // arXiv preprint arXiv:1409.0473. 2014.

[7] Gers Felix A, Schmidhuber Jürgen, Cummins Fred. Learning to forget: Continual prediction with LSTM. 1999.

[8] Graves Alex, Schmidhuber Jurgen. Framewise phoneme classification with bidirectional LSTM and other neural network architectures // Neural Networks. 2005. Т. 18, № 5-6. С. 602–610.

[9] Kim Yunsu, Ney Jiahui Geng Hermann. Improving Unsupervised Word-by-Word Translation with Language Model and Denoising Autoencoder. 2018.

[10] Lample Guillaume, Denoyer Ludovic, Ranzato Marc'Aurelio. Unsupervised machine translation using monolingual corpora only // arXiv preprint arXiv:1711.00043. 2017.

[11] BLEU: a method for automatic evaluation of machine translation / Kishore Papineni, Salim Roukos, Todd Ward [и др.] // Proceedings of the 40th annual meeting on association for computational linguistics / Association for Computational Linguistics. 2002. С. 311–318.

[12] Irvine Ann, Callison-Burch Chris. End-to-end statistical machine translation with zero or small parallel texts // Natural Language Engineering. 2016. Т. 22, № 4. С. 517–548.

[13] Toward statistical machine translation without parallel corpora / Alexandre Klementiev, Ann Irvine, Chris Callison-Burch [и др.] // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics / Association for Computational Linguistics. 2012. С. 130–140.